心理科学进展 2022, Vol. 30, No. 8, 1667–1681
Advances in Psychological Science

https://doi.org/10.3724/SP.J.1042.2022.01667

·国内心理统计方法研究热点回顾(Section of Research Methods)·

编者按:

心理学在中国的发展方兴未艾，而用以支撑心理学研究的各种统计方法也取得了长足的发展。本期刊发温忠麟教授研究团队撰写的一组专栏文章，主题为“新世纪20年国内心理统计方法研究热点回顾”，涵盖了6个热点方向。目的有两个：一是总结新世纪20年国内期刊作者在心理统计方法研究上的贡献，让读者了解国内期刊作者长期致力于“将科研成果写在祖国大地上”，并且在许多热点问题的研究上紧跟国际前沿，部分方向还做出了具有国际先进水平的研究成果。二是关注读者兴趣，就相关议题如何在实际中应用，根据需要通过借鉴发表于英文期刊的研究成果填补空缺，并让读者了解前沿研究进展情况。该组文章在综述的基础上提出了不少创新性的总结和推论。例如，在社科研究领域，存在大量“橘生淮南则为橘，生于淮北则为枳”这类因调节作用导致的不可重复性，因而“心理学研究有可重复性危机”可能是一个伪命题；零假设显著性检验(NHST)已经发展成一套组合方法，各种试图取代NHST的复杂统计方法，不能只是满足于验证是否比NHST更好，而应当看看是否能比这套组合方法更好；如果一个统计模型中的每个变量的合成分数的信度都不小于0.95，使用显变量分析与使用潜变量分析的结果差别不大；等等。希冀该组文章有助于读者恰当应用所论的统计方法，并为心理学的进一步发展带来方法上的启发。

新世纪20年国内假设检验及其关联问题的方法学研究*

温忠麟[1] 谢晋艳[1] 方 杰[2] 王一帆[1]

([1]华南师范大学心理学院/心理应用研究中心，广州 510631)
([2]广东财经大学新发展研究院/应用心理学系，广州 510320)

摘 要 新世纪20年来国内假设检验方法学研究内容可分为如下几类：零假设显著性检验的不足、*p*值的使用问题、心理学研究的可重复性问题、效应量、检验力、等效性检验、其他与假设检验关联的研究。零假设显著性检验已经发展成一套组合流程：为了保证检验力和节省成本，实验研究需要做先验检验力分析预估样本容量，但问卷超过160人在传统统计中就没有必要这样做。当拒绝零假设时，应当结合效应量做出结论。当不拒绝零假设时，需要报告后验检验力；如果效应量中或大而检验力不够高，则可增加被试再行分析，但这一过程应主动披露，报告最后的实际*p*值并对可能犯的第一类错误率做出评估。

关键词 假设检验，*p*值，效应量，检验力，等效性检验
分类号 B841

假设检验是推断统计中的重要内容，通过样本信息来判断对总体参数或总体分布的假设是否可信，包括参数检验和非参数检验(温忠麟，2016；张厚粲，徐建平，2015)。常用的均值差异检验属于参数检验，而正态性检验、独立性检验属于非参数检验。通常报告最多的统计检验结果是根据零假设显著性检验(Null Hypothesis Significance Testing, NHST)做出的。零假设也称为原假设。

国内外不同学科研究者对假设检验都有深入的讨论。在国内，上世纪90年代开始零星出现介绍性质的文章，后面将会提到。新世纪后，相关的研究多了起来，尤其是2003年之后。以中国知网(https://www.cnki.net/)全文数据库为数据源，出版年限设为2001~2020年，关键词包括“假设检验”、“显著性检验”、“显著性水平”、“*p*值”、“效应量”、“效果量”、“检验力”、“检验功效”、“统计功效”和“统计效力”，经筛查得到期刊上发表的有关假

收稿日期：2021-12-29
* 国家自然科学基金项目(32171091)、国家社会科学基金项目(17BTJ035)资助。
通信作者：温忠麟，E-mail: wenzl@scnu.edu.cn

设检验方法学研究论文169篇(不计应用为主的文章)。各学科发文统计情况见表 1，发文较多的学科为：数学与统计 62 篇、医药学 33 篇、心理学29篇，其中心理学期刊上的文章在最近20年快速增长。这些文章可以分为如下几类：对 NHST 的认识，NHST的不足，p值的使用问题，心理学研究的可重复性问题，效应量指标及其大小标准，检验力，等效性检验，其他假设检验关联研究。本文对各类研究进行回顾并做出总结。

效应量(effect seize)和统计检验力(power of statistical test，以下简称检验力)是温忠麟等(2021)总结的新世纪 20 年国内心理统计方法研究 10 个热点之一，本文将其拓展为假设检验及其关联问题，发现数学与统计和医药学期刊发表的论文更多，这与其他热点以心理学期刊论文居多的情况不同，相信这是因为假设检验是统计学的基础，各学科研究者都会感兴趣。

表 1　2001~2020 年国内不同学科假设检验及其关联问题发文数量一览

学科	2001~2005	2006~2010	2011~2015	2016~2020	合计
数学与统计	12	16	19	15	62
医药学	7	13	9	4	33
心理学	2	5	8	14	29
综合性刊物	8	11	3	4	26
工科类	2	4	3	0	9
气象学	1	1	0	0	2
社会学	0	0	2	0	2
经济学	0	0	0	1	1
管理学	0	2	0	0	2
体育学	0	0	1	0	1
教育学	0	0	0	1	1
语言学	1	0	0	0	1
总数	33	52	45	39	169

注：综合性刊物主要包括各高校学报，工科类包括测绘、系统仿真、武器装备试验、军事工程等。

1　零假设显著性检验的认识

研究者们对 NHST 的认识主要分为两个部分，一是深化对假设检验本身的认识，二是澄清应用研究中对假设检验的误解，并提出相应的解决对策。

1.1　深化对假设检验本身的认识

已有研究从不同角度深化了对假设检验的认识，包括假设的确定、两类错误率、区间估计与假设检验、单尾检验、其他假设检验方法等。

杨桂元和刘德志(2012)较为全面地介绍了参数假设检验中的一些概念，包括基本原理、检验的 p 值、两类错误、单尾检验的假设与拒绝域等。吴启富和张玉春(2012)从小概率原理入手，揭示了假设检验的方法论基础，罗荣华和吴锟(2014)则从正态分布下抽样极限误差角度分析了假设检验的相关问题。

对于零假设和备择假设的确定，研究者形成如下共识：零假设与备择假设的地位是非对称的，要根据具体问题谨慎选择合适的零假设，应当将希望其为真的假设(即研究假设)的对立面作为零假设(韩兆洲，魏章进，2005；贺文武，2004；金晓峰，2004；牛莉，2005；杨少华，杨林涛，2009；张凌翔，2006)。

以下研究同时考虑了两类错误。徐浪和马丹(2001)指出零假设的选择要考虑两类错误率。李文华和雷金星(2005)分析了单均值统计检验中的两类错误，认为两类错误不能同时减少。郭宝才和孙利荣(2010)讨论了两类错误率受样本容量的影响情况。房祥忠和陈家鼎(2003)将 Expectation-Maximum 算法运用于假设检验中，不仅分析了两类错误率和临界值，还简化了比较复杂的假设检验问题。张晓敏(2008)基于马氏样本的最优势检验来估计两类错误率，推广了经典的 Neyman-Pearson 基本引理。甘伦知(2011)探讨了对第二类错误的控制，提出需要给出能辨别的最小相对差距，通过选择样本容量可在一定程度上控制两类错误。

就参数的区间估计与假设检验的关系而言，假设检验和区间估计都利用了样本数据的信息来推断总体(樊明智，王芬玲，2006；纪竹荪，2003)，且两者所得的检验结论相同(戴金辉，2019；唐宝珍，2004)。不同之处在于：第一，假设检验是在统计对象的总体参数未知时，通过对总体的部分了解对参数做出某种假设(即零假设 H_0，通常是研究者希望为真的研究假设的对立面)，然后根据样本数据信息判断是否拒绝 H_0。区间估计则是在选定置信水平 $1-\alpha$ 后根据样本数据求得参数可能的范围(区间) (纪竹荪，2003)。第二，假设检验是判断结论是否成立，而区间估计要分析的是范围问

题(樊明智, 王芬玲, 2006)。第三, 区间估计中的置信水平和假设检验中的显著性水平不同(戴金辉, 2019), 即置信水平为 1−α, 显著性水平为 α。第四, 假设检验和置信区间检验中标准误的计算不一样, 假设检验中标准误的估计需要假定 H_0 成立, 而置信区间检验则不需要(何平平, 2004)。

在单尾假设检验的研究中, 钟路(2004)提出当样本统计量恰好位于两个临界值之间时, 应做出由于样本信息不足无法进行统计推断的结论。而彭玉兵(2010)借鉴韦伯−费希纳定律, 提出了一种考虑显著性水平相对增减率的方法, 来解决样本统计量落在接受域与拒绝域的边界时的研究结果问题。另外, 王雪琴(2010)认为均值单尾检验有局限性, 必须进行两次单尾检验(即双侧检验)才能使检验更完善。

一些研究者也提出了新的假设检验方法, 如灰色统计假设检验方法(李勇, 2011, 2012, 2016)、以模糊集合理论为基础建立用隶属度描述的假设检验(林晓辉, 2006a, 2006b; 夏新涛, 王中宇, 2006)、多元模糊数据的假设检验方法(郑文瑞, 丁栋全, 2007)、另类区间估计检验方法等(江海峰, 2009), 但这些方法还极少用于实践。

1.2 澄清假设检验应用中的误解

不同学科都对 NHST 存在一些误解现象, 研究者对此进行了澄清, 如统计显著性与实际显著性的差别(龚凤乾, 2003; 焦璨, 张敏强, 2014), 参数的显著性检验不应该被称为信度检验(黄嘉佑, 2005; 施能 等, 2009), 统计结果显著无法说明实际的效应有多大(陈启山, 2006), 但显示了差异不是由抽样误差造成(何晓东, 2004; 孙红卫 等, 2012), 显著性检验不能够避免两类错误的发生(李世明 等, 2004)。假设检验方法的使用要考虑不同的研究设计和数据类型(田庆丰, 张功员, 2002; 王伟, 2004; 张功员, 2002), 研究者抽样前应当确定合适的样本容量, 并在结果中报告研究的效应量(郭璐, 2016)。另外, 假设检验用于军事工程中要关注两类错误的关联性和样本大小(夏佩伦 等, 2015), 而医学研究的结果要注意统计学意义和临床意义的区别(姚晨, 2007)。当实际应用中出现真值与假设值差异微小的情况时, 为使决策更加客观应当限定样本容量的范围(王雅玲, 2006)。

2 零假设显著性检验的不足与争议

随着对 NHST 认识的加深, 其本身的局限和不足也逐渐显现出来, 进而引发了研究者们的质疑和讨论。

2.1 零假设显著性检验的不足

NHST 的不足主要表现在以下几个方面。第一, 研究结果的显著性无法代表实际显著性(焦璨, 张敏强, 2014; 吴艳, 温忠麟, 2011)。检验结果显著反映了统计上的显著性, 仅表明差异不是由偶然因素造成的(段乃华, 王元佳, 2011), 不同处理组(如实验组与对照组、多种教学方法组等)之间确实存在差异。但统计显著性不等于实际显著性, 读者不应该对统计术语与日常术语做出同样的理解(温忠麟, 2016; 吴艳, 温忠麟, 2011)。第二, 显著性检验的分析过程要求控制第一类错误率(正态情形即显著性水平 α), 通常不考虑第二类错误率(β), 因而对检验力(1−β)的关注也不足。第三, 显著性的检验结果具有不稳定性(焦璨, 张敏强, 2014; 张静, 2012), 检验研究假设的过程中选择不同的统计量可能会得出不一致的结论(黄闪闪, 高瑞华, 2014)。第四, 零假设的选择可能是任意指派的(黄闪闪, 高瑞华, 2014), 导致其设定主观性太强(罗晓娟, 2011)。此外, 假设检验的不足还有不能同时考察多个研究假设(张静, 2012)和依赖样本容量(焦璨, 张敏强, 2014)。

2.2 零假设显著性检验的争议

NHST 自身存在的不足引发了学术界的讨论, 有研究者认为 NHST 存在逻辑漏洞, 应用价值不大(辛自强, 2010)。但温忠麟和吴艳(2010)回应了这一说法, 认为 NHST 是可用的, 因为显著性已经告诉了研究者根据样本推断的结果多大程度是由抽样造成的。他们还介绍了国外争议的情况, 结论是 NHST 可以继续使用, 但仅仅报告 NHST 结果是不够的。

从假设检验遭受的争议来看, 吕小康(2012)在分析了假设检验思想的提出者 Fisher 与 Neyman-Pearson 在统计模型方法论、两类错误率、显著性水平等方面存在的分歧后, 认为 NHST 存在的争议是心理统计的教育模式造成的, 是对于统计推论背后的思想了解匮乏导致的。而仲晓波等(2008)的研究回应了心理学中对假设检验的批评, 还提出对于绝大部分心理学实验来说, 数据分析适合

采用频率学派的方法，建议报告效应量及其置信区间作为对传统假设检验的改进和补充(仲晓波, 2010a, 2010b, 2016)。

尽管 NHST 饱受争议和批评，但它依然是社会科学实证研究中常用的假设检验方法，因为它满足了研究者追求知识的客观性与确定性，以及将自身学科发展成为一门科学分支的期望，从这个意义上来说，NHST 已从普通的统计工具上升为了一种科学范式(吕小康, 2014)。

3 *p* 值的使用问题

p 值是零假设 H_0 为真时，样本数据结果或更极端结果出现的概率(简记为 P(样本|H_0))。但 *p* 值的含义容易引起误解，下面文献指出了在应用中对 *p* 值的各种误解。误解 1: *p* 值是给定样本数据条件下 H_0 的正确概率，衡量了零假设成立的可能性(傅军和, 2009; 吕佳, 乔克林, 2010; 孙红卫 等, 2012)，实际上是将条件概率 P(样本|H_0)误解为条件概率 P(H_0|样本)了。误解 2: *p* 值表示了实际效应差异大小(王伟, 2004; 张弓, 肖景榕, 2006)，*p* 值越小反映组间差异越大(陈薇 等, 2011; Lu & Belitskaya-Levy, 2015; 孙红卫 等, 2012)。误解 3: *p* 值越小表明重复实验中出现有统计学意义结果的可能性越大(孙红卫 等, 2012)。误解 4: *p* 值越大反映了支持零假设的证据越强(陈薇 等, 2011)。此外，*p* 值还有一个缺点，即当样本容量很大时，总能得到很小的 *p* 值(吕佳, 乔克林, 2010; 孙红卫 等, 2012)。因此，应用工作者需要正确理解 *p* 值，并报告具体的 *p* 值(陈薇 等, 2011; 李康, 2005; 张弓, 肖景榕, 2006)。

2016 年美国统计协会发表了《关于统计显著性与 *p* 值》的官方声明，提出了 6 条正确使用 *p* 值的准则(Wasserstein & Lazar, 2016)，这份声明再次引起了国内科研工作者对 *p* 值的广泛讨论。不同领域的研究者从各自的研究视角讨论了 *p* 值的使用情况，并提出了一些补充和改进方法。郝丽等(2016)建议基于 *p* 值的推理要有完整的研究报告和透明的研究过程，余红梅(2017)提出要报告准确的 *p* 值并综合使用假设检验，在计算 *p* 值前给出显著性水平 α (沈光辉 等, 2019)，并引入置信区间(程开明, 李泗娥, 2019; 金辉, 邹莉玲, 2017; 余红梅, 2017)、效应量(程开明, 李泗娥, 2019; 沈光辉 等, 2019; 宋爽, 曹一鸣, 2019; 余红梅, 2017)、检验力(程开明, 李泗娥, 2019)、贝叶斯因子(程开明, 李泗娥, 2019; 余红梅, 2017)、错误发现率(程开明, 李泗娥, 2019)等指标作为 *p* 值的替代和补充，多进行重复实验(程开明, 李泗娥, 2019)，也可使用元分析方法(余红梅, 2017)。

4 心理学研究的可重复性问题

2015 年开放科学协作组的科学家团队在 *Science* 上发表了一项研究“Estimating the reproducibility of psychological science”，重复了刊登在心理学顶级期刊上的 100 项研究，结果只有 36%的实验结果得到重现(Open Science Collaboration, 2015)。这一研究受到了心理学及整个社会科学领域的关注，展开了关于心理学研究可重复性问题的探讨。

不少人将心理学研究可重复性危机归因于 NHST 体系(胡传鹏 等, 2016; 刘佳 等, 2018; 骆大森, 2017)。具体地，骆大森(2017)得出心理学研究可重复性危机有两大来源，一个是传统虚无假设显著性检验体系的制约，另一个是非统计学因素，包括人为偏误、发表偏见和可疑研究操作等。仲晓波(2015)认为是过多的额外变量导致了心理学实验研究的可重复性较低。聂丹丹等(2016)认为统计显著性检验的不确定性、样本和检验力问题、统计方法和模型误用、实验设计灵活和选择性报告是可重复性问题的原因。刘佳等(2018)提出研究人员的偏差性操作是影响可重复性的重要原因。胡传鹏等(2016)认为心理学研究的可重复性问题是因为发表的研究假阳性过高，而更深层的原因却是出版偏见和过度依赖虚无假设。

就如何提高研究的可重复性，研究者一方面建议心理学的研究结果要报告效应量及其置信区间作为检验结果的补充(吴艳, 温忠麟, 2011; 仲晓波, 2010b, 2015, 2016)，另一方面提出使用贝叶斯学派的统计检验方法作为 NHST 的替代或补充，计算贝叶斯因子来做出统计决策(胡传鹏 等, 2018; 吴凡 等, 2018)。但目前贝叶斯因子的应用还很有限，分析软件也少(许岳培 等, 印刷中)。

5 效应量指标及其大小标准

效应量，也称为效果量，是衡量实验处理效应的指标。它不仅反映了统计检验效应的大小，也反映了两个总体受某事物影响后的差异程度

(胡竹菁, 2010)。效应量能够区分统计显著性和实际显著性, 估计检验力, 并通过元分析方法比较前人的研究结果(郑昊敏 等, 2011)。张力为和祁国鹰(1998)率先介绍并在运动心理学研究中使用了效应量。国内学者对效应量的研究主要分为以下几类: 一是介绍科学研究报告中常用的几种效应量, 并采用具体例子对效应量的计算方法和使用标准进行阐述, 二是对多种类型的效应量指标进行了归纳和分类, 以便读者在不同的条件下选择和报告合适的效应量; 三是探讨了什么样的统计量可以作为效应量的指标, 分析了效应量指标应具备的性质。

5.1　常用的效应量指标

权朝鲁(2003)最先介绍了心理学研究中的几种效应量及其评价标准, 即 d, r_{pb}^2, η^2 和 ω^2 的测定方法。而温煦(2011)也描述了体育科研中常用的效应量指标及其标准, 即 d, η^2, η_p^2 和 φ。胡竹菁等人有系列研究详述了 Z 检验、t 检验、F 检验和 χ^2 检验下的效应量指标(d, η^2, η_p^2, φ 和 Cramer's_V)及其计算方法和评价标准(胡竹菁, 2010; 胡竹菁, 戴海琦, 2011, 2017)。刘铁川等(2019)介绍了一种方差分析效应量的新指标——广义 eta 方, 可以同时考虑操作因素和个体差异, 实现跨研究设计效应量的可比性, 但在国内的应用不多, 并且无法计算置信区间。沈光辉等(2019)也介绍了教育研究中均值 t 检验、方差分析(F 检验)、回归系数检验、相关系数检验和 χ^2 检验的常用效应量指标。另外, 李海峰和姜小峰(2014)还介绍了病例对照研究中用比值比反映的效应量 OR 值和 Q 值。

续志琦和辛自强(2018)分析了单被试实验的 5 种基于非重叠法的效应量指标(即扩速线指数、提高率差异、非重叠对占比、控制基线趋势的非重叠 Tau 值和非重叠数据占比), 并结合实际例子进行了阐述, 最后提出了非重叠法效应量的选择和评价标准: 不仅要根据实验数据特征选择合适的效应量指标, 还需要考虑效应量指标的鉴别力、精度和检验力等因素。

方杰等(2012)介绍了 4 种中介效应的效应量指标, 即 P_M, R_M, R^2_{med} 和 κ^2, 建议使用 R^2_{med} 和 κ^2 指标及其置信区间。而温忠麟等(2016)明确指出 R^2 型指标和 κ^2 都缺乏单调性, Wen 和 Fan (2015)已经证明了把 ab 的最大可能值作为 κ^2 的分母是错误的, 终结了 κ^2 这个在国际上曾经流行的中介效应量的合法性, 建议同时报告多个中介效应量指标的原始估计和标准化估计。

5.2　效应量指标分类

研究者对目前存在的多种类型的效应量指标进行了分析和总结, 详见表 2。郑昊敏等(2011)将效应量划分为差异类、相关类和组重叠类三种类型, 卢谢峰等(2011)则将效应量区分为标准差异型和关联强度型两类。焦璨和张敏强(2014)根据汤普森的划分标准, 将效应量指标区分为三类, 即标准化平均数差异效应量, 未调校的考虑方差的效应量和调校的考虑方差的效应量。蒲显伟(2016)认为效应量可分为组间差异类(d 类)和相关系数类(r 类)两类, 但未具体介绍对应的效应量指标, 而是按照参数检验和非参数检验的不同方法详细介绍了效应量。总的来看, 效应量的分类较为类似, 一致的意见是将效应量指标分为差异类和其他类别。

表 2　效应量指标分类

国内文献	效应量分类	对应的指标
郑昊敏等 (2011)	差异类	Cohen 的 d, Glass 的 Δ, Hedge 的 g
	相关类	r、r_{pb}、r_b、$r_{\text{equivalent}}$、φ 及 Cramer 的 V 系数等基于 χ^2 统计量的相关系数等; 方差比 f^2, R^2, η^2, ω^2, ε^2; 以及 r_{alerting}, $r_{\text{effectsize}}$, r_{contrast} 等
	组重叠	Improvement-Over-Chance index(I 效应量)
卢谢峰等 (2011)	标准差异型	d, Δ, g, g_D, $g_{\text{corrected}}$
	关联强度型(非平方尺度)	φ, V, r, r_{pb}
	关联强度型(平方尺度)	η^2, η^2_{partial}, ω^2, R^2, R^2_{partial}, R^2_{adjusted}
焦璨和张敏强 (2014)	标准化平均数差异效应量	Hedges 的 g, Cohen 的 d
	未调校的考虑方差的效应量	R^2, η^2
	调校的考虑方差的效应量	Ezekiel 的 R^{2*}, ω^2
蒲显伟(2016)	组间差异类(d 类)、相关系数类(r 类)	

5.3 效应量指标的性质

效应量表示了研究结果的实际显著性，是元分析和检验力分析不可缺少的参数。那究竟哪些统计量适合作为效应量的指标呢？温忠麟等(2016)提出了效应量指标应当具有的一些性质：(1)与测量单位无关，而得到与测量单位无关的效应量有两种方式，一是标准化效应，二是将效应量定义为一种比例。(2)相对于效应而言具有单调性，即其他条件不变的情况下，研究中感兴趣的效应(绝对值)越大，效应量(绝对值)也应该越大。(3)不受样本容量的影响，也就是效应量不会随样本容量的增大而系统变大。其他还可考虑的性质有非负性、有界性和正规性。

5.4 效应量及其评价标准总结

美国心理协会写作手册从 1994 年起要求研究者报告心理学实验的效应量和检验力，我国心理学重要期刊则从 2013 年开始明确要求报告效应量，报告效应量已成为心理学研究论文发表的标准之一。而效应量作为假设检验的补充，不少研究者还建议增加报告效应量的置信区间(卢谢峰 等，2011；吴艳，温忠麟，2011；仲晓波，2010b，2015，2016)，因为这样有利于比较不同研究间的误差大小，提供更丰富的信息，也能帮助研究者得出正确的结论(王珺 等，2019)。为增强应用工作者对效应量置信区间的理解和应用，王珺等(2019)以 t 检验中 d 和方差分析中的 η^2 为例，展示了效应量置信区间的计算公式和软件实现过程。

在不同的研究条件和实验设计下，可供选择的效应量指标很多。综合国内外已有的效应量研究结果，表 3 总结了常见统计方法的常用效应量指标及其评价标准。

当然，提高实验研究的效应量更应该通过完善研究设计和减少实验误差来实现。效应量的评价标准并不存在唯一准则，需要结合研究主题、理论背景、研究设计类型、实验控制过程等多种因素来确定(卢谢峰 等，2011)，也可以参考元分析报告或者同类研究的结果。

6 检验力

检验力，有的文献也称为统计功效、检验效能、检验功效、统计效力等，是 H_0 为假时正确拒绝 H_0 的概率。金炳陶和马承霈(1992)率先介绍了检验力。国内有关检验力的研究可分为统计方法的检验力分析和研究效应的检验力分析。统计方法的检验力指的是某种统计方法能有多大的可能性检测到真实存在的差异，而研究效应的检验力指的是研究者感兴趣的某些研究效应被不同研究重复发现的可能性。影响检验力的因素有效应量、样本容量和显著性水平，如果保持其他条件不变，检验力会随效应量、样本容量和显著性水平的增大而提高(温忠麟，2016；吴艳，温忠麟，2011)。

6.1 统计方法的检验力

统计方法的检验力分析主要集中于统计学和医药学两个领域。统计学的研究探讨了重复测量

表 3 研究报告中常见效应量及其评价标准

统计分析方法	效应量	评价标准
t 检验	$d=\frac{(\bar{X}_1-\bar{X}_2)}{S_{\text{pooled}}}$	0.2 为小，0.5 为中，0.8 为大
相关分析	皮尔逊相关系数	0.1 为小，0.3 为中，0.5 为大
方差分析	$\eta^2=\frac{SS_{\text{effect}}}{SS_{\text{total}}}$	0.01 为小，0.06 为中，0.14 为大
回归分析	R^2	0.02 为小，0.13 为中，0.26 为大
中介效应分析	$P_M=\frac{ab}{c}$ (ab 和 c' 同号)	总效应按相关系数标准，中介效应占总效应的 20%以上
调节效应分析	加入调节项后，回归方程的 R^2 变化值(ΔR^2)	调节项额外解释因变量比例不低于 2%

注：t 检验、相关分析、方差分析效应量评价标准参见 Cohen (1988)；回归分析则按其中的效应量 f^2 的标准(0.02 为小，0.15 为中，0.35 为大)换算得到，不过，如果自变量只有一个，应将相关系数的平方作为效应量。有文献(如蒲显伟，2016)指出，这些切分点(如 d = 0.2)可看成是效应量(小)的区间下限。中介效应和调节效应分析未见到有标准，但对于传统的中介模型，总效应应当达到 0.2 左右、中介效应占比超过 20%才有意义；对于调节效应，有元分析发现调节项额外解释因变量比例通常为 3%~8% (Champoux & Peters, 1987)，应当不低于 2%才有意义(温忠麟，叶宝娟，2014)。

试验模型(侯紫燕, 廖靖宇, 2007)和多元重复测量试验模型(侯紫燕, 原新凤, 2007)的似然比检验的统计功效, 功效函数的蒙特卡洛模拟(张建侠, 鞠银, 2012), 双幂变换下正态线性回归模型的功效函数分析(丘甜 等, 2017), 三种非参数检验方法(卡方检验、Mann-Whitney 和 Kolmogorov-Smirnov 检验)的功效分析(刘遵雄, 2018)。而医药学领域研究了 logistic 回归(刘韵源 等, 2001; 王慧 等, 2019)、p 值分布的百分位数(丁守銮 等, 2004)、变量变换(魏杰, 吴学森, 2006)、两组 t 检验与秩和检验(颜杰 等, 2004)、非参数检验方法(Wilcoxon、Kruskal-Wallis、Median 和 Kolmogorov-Smirnov)的检验功效(曾艳 等, 2011; 张超 等, 2008)。

6.2　研究效应的检验力

就研究效应的检验力而言, 吴艳和温忠麟(2011)认为检验力分析可以分为先验检验力分析(prior power analysis)和后验检验力分析(post-hoc power analysis), 可用于检验力分析的软件有 nQuery Advisor Release、Sample Power、G*Power、UnifyPow 和 PASS 等。先验检验力分析侧重于实验前计算研究所需的样本容量, 后验检验力分析则关注数据收集和分析后的实验效应的检验力有多高。下面分别来看一下这两种检验力分析的相关研究。

6.2.1　先验检验力分析

在研究开展前根据给定的显著性水平、效应量和检验力, 估计研究所需的样本容量可节省实际研究的成本, 这也就是进行了先验检验力的分析。医药学领域中陈平雁(2015)介绍了临床试验中常用的样本容量估计方法和专业软件操作, 万霞和刘建平(2007)推导了截面研究、观察性研究的样本容量计算公式, 其他研究者也提供了 logistic 回归(刘韵源 等, 2001; 王慧 等, 2019)、两样本均数比较、两样本率比较和分层设计的样本容量计算公式(万霞 等, 2007)。其他学科领域也有部分先验检验力的研究, 如方差检验(郭文, 2012)、方差区间估计和假设检验(耿修林, 2008)、正态总体期望研究中 Bayes 假设检验(贾旭山, 金振中, 2012)、泊松分布参数的序贯概率比检验(赵盼, 宋学力, 2016)、Bayes 最小样本容量截尾值序贯检验(胡思贵, 王红蕾, 2019)等方法中的样本容量计算。

先验检验力分析的目的是为了确定研究所需的样本容量(即被试人数), 这在被试不易得到或者实验成本较高的时候(如医学实验、使用高级设备的心理与脑实验)很有必要。为了方便研究者, 我们采用 GPower 3.1.9.7 计算了常用的检验方法所需的被试人数(见表 4)。设定检验力为 0.8, 显著性水平为 0.05 和 0.01, 效应量为小、中和大三种, 给出了相应方法在双侧检验时需要的被试总人数。单侧检验时, 需要的被试比双侧检验的要少。

对于常见的统计方法, 从表 4 中可以总结出两点: 第一, 即使是小效应量, 在 0.05 显著性水平上, 估算的被试人数都不超过 1 千; 第二, 注意到对于通常的研究, 效应量小的时候, 即使效应显著意义也不大(温忠麟 等, 2016); 而效应量中或大的时候, 在 0.05 显著性水平上, 估算的被试人数不超过 160, 所以当被试人数超过 160 时, 不需要做检验力分析去确定被试人数。

表 4　常用检验方法的被试人数估算

统计方法	效应量(α = 0.05)			效应量(α = 0.01)		
	小	中	大	小	中	大
配对样本 t 检验(或单样本 t 检验)	199	34	15	296	51	22
独立样本 t 检验($n_2/n_1 = 1$)	788	128	52	1172	192	78
独立样本 t 检验($n_2/n_1 = 0.5$)	591 和 295	96 和 48	39 和 19	879 和 439	143 和 71	57 和 29
单因素方差分析(被试间, 3 水平)	969	159	66	1395	228	93
两因素方差分析(被试间 2×2)	787	128	52	1172	191	77
三因素方差分析(被试间 2×2×3)	967	158	64	1393	227	92
单因素方差分析(被试内, 3 水平)	163	28	12	234	40	17
两因素方差分析(被试内 2×2)	138	24	10	196	33	14
相关分析(2 个连续变量)	782	84	29	1163	125	42
回归分析(2 个自变量)	485	68	31	699	98	45

注: 按检验力为 1−β = 0.8 估算的被试总人数。效应量大小标准见表 3。

6.2.2 后验检验力分析

胡竹菁(2010)给出了两独立样本平均数差异显著性检验的后验检验力估计方法，根据样本计算的 Z (或 t)值和 α 水平临界值，确定可能犯的第二类错误率，进而求得检验力 1−β 的概率。胡竹菁和戴海琦(2011)给出了方差分析的后验检验力的计算步骤。赵礼和王晖(2019)详细描述了后验检验力的影响因素和基本分析流程，并演示了如何用 Optimal Design 软件分析多层模型的检验力。而其他讨论后验检验力分析的论文也见于医药学(钱俊，陈平雁，2005；吴迪 等，2007；姚嵩坡 等，2010)和管理学(陈功兴，容迪，2010；林丹明 等，2008)。但从逻辑上说，只有检验结果不显著时，才需要计算并报告后验检验力。因为检验结果显著时，只可能犯第一类错误，而报告检验力相当于报告第二类错误率(后验检验力 = 1 − 第二类错误率)。

7 等效性检验

以差异检验(包括效应是否为零、均值是否相等)为例，通常的零假设是无差假设，而备择假设是想要验证有效应(如效应不是零、均值不相等)的假设。当拒绝零假设的时候，犯错误的概率是 α (通常是 0.05)，不仅明确已知，而且已经受控。但如果想要验证的就是等效(效应为零、均值相等)的呢？如果还将无差假设作为零假设，接受零假设的时候，犯错误的概率(第二类错误率)不仅需要后验检验力分析，而且往往都比较大(例如超过 0.2)。一种解决的办法是等效性检验(equivalence testing)：借鉴效应量的做法，效应要达到或超过一个界值才算有效，并将其作为零假设，这样就把希望为真的等效性假设放在备择假设的位置。

等效性检验是 NHST 的延伸，它用来检验两个总体的差异是否在某范围之内(王静，胡镜清，2011)。等效性检验的零假设(H_0: $|\mu_1-\mu_2|\geq c$)可理解为：实验组的效应 μ_1 与对照组的效应 μ_2 的差异超过了等效的界值 c (c 是一个小的正数)。备择假设(H_1: $|\mu_1-\mu_2|<c$)可理解为：实验组的效应 μ_1 与对照组的效应 μ_2 的差异在等效范围$(-c, c)$内。等效性检验需要进行两次单侧的 NHST，一次单侧检验的零假设是(H_0: $\mu_1-\mu_2\geq c$)，拒绝零假设说明实验组非优效；另一次单侧检验的零假设是(H_0: $\mu_1-\mu_2\leq -c$)，拒绝零假设说明实验组非劣效。只有两次单侧检验的 p 值都小于显著性水平 α，才能得出实验组和对照组的效应等效的结论(王静，胡镜清，2011；于莉莉 等，2005)。

等效性检验和 NHST 有如下区别。第一，假设的差异。等效性检验的假设都是围绕实验组和控制组的效应之差($\mu_1-\mu_2$)与等效的界值 c 的关系设定的，而 NHST 的假设都是围绕 $\mu_1-\mu_2$ 和 0 的关系设定的，这是两类检验的本质差异(王静，胡镜清，2011)。因此，NHST 仅有统计学上的意义，而等效性检验则关注临床上或实践中有没有效应。第二，检验的目的有差异。等效性检验的目的是验证实验组与对照组的效应是否足够接近(即等效)，而 NHST 的目的是检验实验组与对照组的效应之差是否足够大，大到能在统计上的识别出来。在这个意义上，等效性检验和 NHST 的作用刚好相反。第三，NHST 的“差异有统计学意义” (即 $p<$ α)也有可能实验组和控制组的效应是等效的，NHST 的“差异无统计学意义” (即 $p>$ α)并不表示实验组和控制组的效应一定等效(于莉莉 等，2005)。

此外，安胜利的系列研究也分析了显著性检验和等效性检验的联系与区别(安胜利，2007a，2007b；安胜利，陈平雁，2007)，并给出了不同条件下基于 p 值进行等效性判定的标准。其他的讨论还有非劣效性试验数据的假设检验(李路路 等，2014；刘玉秀 等，2008)、生物等效性(一种等效性检验)研究的受试者数量和事后统计功效等(代骏豪，郑强，2017；贺江南 等，2009)。

8 假设检验的其他关联研究

其他与假设检验关联的研究主要为 NHST 与贝叶斯假设检验的比较、具体统计方法的假设检验问题。

对于贝叶斯假设检验和 NHST 的比较，尹玉良等(2011)发现频率学派和贝叶斯学派在正态模型单边假设检验中得到的证据具有一致性。但更多研究讨论了贝叶斯检验比显著性检验的优势：第一，可利用合理的先验信息和抽样信息减少决策损失(李楚进，万建平，2015)，但先验信息的选择通常是一个难点；第二，避免显著性检验的主观性问题(黄闪闪，高瑞华，2014；李楚进，万建平，2015)；第三，同时考虑 H_0 和 H_1 并可以用来支持 H_0，且可监控证据强度的变化(胡传鹏 等，

2018); 第四, 揭示备择假设与虚无假设成立可能性的高低(吴凡　等, 2018)。

对于具体统计方法的假设检验关联研究涵盖内容较多(详见表 5), 在此不一一叙述。

9　总结与讨论

9.1　零假设显著性检验还可继续使用

NHST 从数理统计应用到包括心理学在内的各个学科, 经历了从认识、使用、误解、澄清、质疑、不断提出改进和替代方法的一系列过程。有关 NHST 理论和方法的研究多采用公式推导、数据模拟和实例验证的方式, 而且多集中于数学与统计、医药学、工科类的研究领域, 而其他学科领域多采用文献综述的方式介绍和澄清假设检验的相关内容。

NHST 还可以继续使用, 但需要有正确的认识: 首先, 尽管 NHST 的不足和质疑引发了激烈讨论, 但它的地位依然稳固, 因为它已表明了显著性的研究结果很不可能由抽样波动造成。第二, 显著性检验的 p 值表示概率 P(样本|H_0), 而不是 P(H_0|样本), Anderson (2020)采用模拟研究分析了在不同条件下两者之间的差异。第三, 在报告显著性结果时, 建议报告准确的 p 值, 以对第一类错误率有更精确的评估。

9.2　零假设显著性检验已经发展成一套组合拳

虽然 NHST 仍可以继续使用, 但不仅要报告统计检验结果, 还要报告效应量(如果显著)或检验力(如果不显著), NHST 的流程如下(见图 1):

第一, 采样前要进行先验检验力分析, 计算出合适的样本容量。但对于常见的统计分析(如线性回归和方差分析等), 问卷研究被试超过 160 人通常不必做先验检验力分析。

第二, 收集数据, 分析并报告参数的 NHST 检验结果和置信区间。

第三, 如果统计显著(此时只可能犯第一类错误), 计算并报告效应量, 根据效应量大小做出结论。

第四, 如果统计不显著(此时只可能会犯第二

表 5　具体统计方法的假设检验关联研究

方法	假设检验的内容
均值比较	贝叶斯样本均值假设检验(林晓辉, 2001), 样本量与方差对 t 检验和 u 检验的影响(金晓峰, 2004), 两组均值比较似然比检验(邓文丽, 2003), 多元总体均值差异显著性检验(田晓明, 傅珏生, 2005), 多维正态总体零均值假设检验(李荣华, 徐九韵, 2001), 正态总体均值与标准差比的置信区间检验(何春, 2011), 两总体均值半参数假检验(万树文, 方芳, 2012), 正态总体均值区间估计和假设检验的 R 函数(张应应, 魏毅, 2014)
方差分析	方差分类模型的假设检验(王石青, 史慧娟, 2007), 广义 p-值法在异方差时优于广义 F-检验(扈慧敏, 徐兴忠, 2007), 方差的区间估计和假设检验的 R 函数(张应应, 魏毅, 2014), 基于最小广义特征值的两因素多元方差分析检验(江忠伟, 郭新颖, 2018), 引入虚拟变量的单因素方差分析(傅莺莺　等, 2019)
相关分析	相关系数显著性检验的几何意义(姚菊香　等, 2007), 独立总体和相关总体的相关系数假设检验(江梅, 2010), 小样本 Kendall τ 相关系数显著性检验(胡春健, 2013)
不同分布	二项分布假设检验平均试验数公式(孙晓峰, 赵喜春, 2003), 二项分布贝叶斯假设检验方法(贾旭山, 金振中, 2008), 两个样本正态分布密度比的假设检验方法(牟唯嫣, 熊世峰, 2009), 两均匀分布总体区间长度比的区间估计和假设检验方法(郑发美, 2009), 混合 Pareto 分布的假设检验问题(刘媚, 2011), Lomax 分布参数的区间估计和假设检验问题(龙兵, 2014), 二维连续型分布密度函数假设检验方法(张凤宽, 2012), 总体非正态时逼近统计量分布的数据的假设检验(魏艳华　等, 2018)
不同模型	坐标转换模型中尺度参数假设检验模型(徐天河, 杨元喜, 2001), 线性半参数模型非参数假设检验(丁士俊, 姜卫平, 2014), 线性混料模型的假设检验问题(黄秀秀, 张崇岐, 2014), 序约束下带有协变量的序贯 k-out-of-n 模型的假设检验问题(杜宇静, 姜丽萍, 2016), 含方程误差的重复测量误差模型参数的假设检验方法(王雅慧, 曹春正, 2016), 非平稳二元选择模型的显著性检验方法(徐鹏　等, 2016), 双幂变换下正态线性回归模型参数的假设检验问题(丘甜　等, 2017), 某一类随机偏微分方程极大似然估计的假设检验问题(王潇文, 吕艳, 2020)
其他	和分布统计量用于小样本离散型多总体的假设检验问题(潘高田　等, 2001), 假设检验的相对稳定性(林路, 张润楚, 2001), 变异的假设检验(李胜联　等, 2006; 苟鹏程　等, 2006), K 个单参数指数总体相等的假设检验方法(宋立新, 张平, 2009), 指数族下参数双侧检验的 p-值(谢田法, 吴启光, 2011), 对应分析应用中的假设检验问题(李克均　等, 2008), 多重假设检验的参数估计问题(刘遵雄, 田珊珊, 2017), 大维数据总体协方差是否等于 Σ_0、$\sigma^2\Sigma_0$ 的假设检验问题(王晓波, 李会琼, 2017), 权数可靠性的假设检验范式(谢忠秋, 2018)

类错误), 计算效应量, 当效应量小时接受零假设; 当效应量中等或大时, 则需进行后验检验力分析: 如果检验力高, 则接受零假设; 如果检验力不到80%, 则可增加样本容量重新分析结果并做出判断。但增加样本容量的这一过程应主动披露, 报告最后的实际 p 值并对可能犯的第一类错误率做出评估, 因为中途增加被试会导致第一类错误率的增加。

Sagarin 等(2014)提出了 $p_{augmented}$ 指标来衡量数据增加带来的一类错误率的膨胀程度。$p_{augmented}$ 的计算基于初始样本容量(N_1)、增加的样本容量(N_2)、统计显著性的临界值(p_{crit}, 通常设置为 0.05)和最终组合数据集中的 p 值($p_{combined}$), 其论文中也提供了相应的 R 脚本和 Excel 计算表(http://www.paugmented.com)。也有一些学者提出了独立分段程序(independent segments procedure)、序列概率比 t 检验(sequential probability ratio t test)等方法(Miller & Ulrich, 2021; Schnuerch & Erdfelder, 2020), 控制统计决策错误概率, 提高研究效率。但在实际应用中可以简单化, 如果最终结果在 0.05 水平上显著, 那么第一类错误率基本上在 0.08 以下; 换一个角度说, 如果最后得到的 p 值小于 0.01, 那么第一类错误率基本上不会超过 0.05。

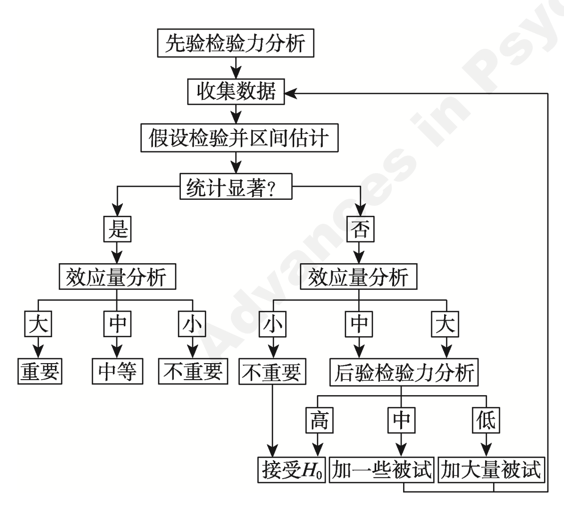


图 1 零假设显著性检验的统计分析流程图 (吴艳, 温忠麟, 2011)

图 1 所示的 NHST 流程其实是一套组合拳, 既要看显著性结果, 也要看效应量大小甚至检验力, 综合做出推断。一方面, 这套组合拳可以避免单纯依靠显著性($p < 0.05$)可能引起的选择性报告数据、p 值操纵等现象, 避免得到假阳性结果(显著但效应量低), 同时避免检验力低导致的假阴性(不显著但有不低的效应量和实际意义); 另一方面, 各种试图取代 NHST 的复杂统计方法, 不能只是满足于验证比 NHST 优胜, 而应当看看是否能比上述的组合拳优胜。遗憾的是, 目前各种试图取代 NHST 的复杂统计方法(如贝叶斯因子法)都只显示与单纯的 NHST 结果比较有优势, 而未有考虑与上述的组合拳比较有优势, 因而未能确定替代方法是否更好。不过, 如果作为补充方法使用是可取的, 可以提供多一点信息。

9.3 深究一下“可重复性”问题

心理学研究的可重复性问题也部分归因于 NHST 这一检验模式。但是, 可重复性问题需要严格地界定, 否则“可重复性”在社科研究领域可能是伪命题。在社科研究领域, 既有大量的“种瓜得瓜、种豆得豆”那样的可重复性主效应, 也有大量“橘生淮南则为橘, 生于淮北则为枳”那样的因调节作用导致的不可重复性。种族、文化背景、年龄、地域、时间等等都可能是调节变量, 使得研究效应时强时弱。当一项研究不能重复时, 虽然有可能是操作不严谨、方法不当造成的, 但也可能是调节作用造成的, 重复研究的时候毕竟是时过境迁, 不能简单看是否能重复去评判一项研究的科学性。

9.4 相关议题的研究拓展

结构方程中的模型拟合检验、测量不变性检验都是希望得到不显著的结果, 等效性检验的思想很适合这类检验。已有研究将等效性检验拓展到结构方程模型评价(Yuan & Chan, 2016; Yuan et al., 2016; 王阳 等, 2020), 做法还是设定适当的“等效”界值并改变零假设。

检验力方面的拓展是针对传统统计以外的模型进行检验力分析。例如, 针对中介效应模型的检验力分析(Schoemann et al., 2017; Zhang, 2014), 针对结构方程模型的检验力分析(Wang & Rhemtulla, 2021)。

效应量的拓展是在传统统计以外的模型中, 利用方差分解提出新的 R^2-型效应量。例如, Rights 和 Sterba (2018)提出单层和多层回归混合模型(regression mixture model, 回归混合模型允许截距和斜率因潜在类别而异)的 12 种 R^2-效应量。Rights 和 Sterba (2019)将因变量的方差进行分解,

提出多层线性模型的 12 种 R^2-效应量。Liu 和 Yuan (2021)将因变量的方差进行分解，提出调节效应的 4 种 R^2-效应量。Liu 等(in press)将中介效应的方差进行分解，提出有调节的中介效应的效应量 φ，即中介效应的方差中有多少能被调节变量解释。刘红云等(2021)将自变量对因变量的效应的方差进行分解，提出了有中介的调节效应的效应量 ϕ，即自变量对因变量的效应的方差中，能被有中介的调节效应解释的比例。

参考文献

安胜利. (2007a). 假设检验应用中的常见问题及改进方法. *南方医科大学学报, 27*(3), 382–389.

安胜利. (2007b). 用传统显著性检验方法进行等效性检验的规律研究. *中国药房, 18*(26), 2077–2080.

安胜利, 陈平雁. (2007). 等效性检验与差异性检验的区别及其模拟验证. *中国卫生统计, 24*(3), 226–228.

陈功兴, 容迪. (2010). 统计效力和效应量的估计方法与应用. *企业科技与发展*, (22), 132–133.

陈平雁. (2015). 临床试验中样本量确定的统计学考虑. *中国卫生统计, 32*(4), 727–733.

陈启山. (2006). 心理学研究中应用统计方法应注意的几个问题. *心理与行为研究, 4*(3), 200–206.

陈薇, 郑国华, 刘建平. (2011). 正确理解与阴性结果试验相关的统计学概念. *中西医结合学报, 9*(5), 487–490.

程开明, 李泗娥. (2019). 科学研究中的 P 值: 误解、操纵及改进. *数量经济技术经济研究, 36*(7), 117–136.

戴金辉. (2019). 区间估计与参数假设检验的比较. *统计与决策, 35*(9), 72–74.

代骏豪, 郑强. (2017). 生物等效性研究中的受试者例数确定和事后统计功效. *中国新药杂志, 26*(24), 2892–2897.

邓文丽. (2003). 重复测量中两组均值是否相等的假设检验. *应用概率统计*, (2), 198–202.

丁士俊, 姜卫平. (2014). 线性半参数模型非参数假设检验理论和方法. *武汉大学学报(信息科学版), 39*(12), 1467–1471.

丁守銮, 王洁贞, 孙秀彬, 傅传喜, 郭冬梅. (2004). 单样本和两样本单侧 Z 检验 P 值的理论分布及应用. *中国卫生统计, 21*(3), 28–32.

杜宇静, 姜丽萍. (2016). 序贯 k-out-of-n 系统在序约束下参数的假设检验. *吉林大学学报(理学版), 54*(3), 487–492.

段乃华, 王元佳. (2011). 精神医学中的生物统计(1) 显著性检验与可信区间. *上海精神医学, 23*(1), 60–63.

樊明智, 王芬玲. (2006). 区间估计与假设检验. *统计与决策*, (12), 141–143.

方杰, 张敏强, 邱皓政. (2012). 中介效应的检验方法和效果量测量: 回顾与展望. *心理发展与教育, 28*(1), 105–111.

房祥忠, 陈家鼎. (2003). EM 算法在假设检验中的应用. *中国科学(A 辑: 数学), 33*(2), 180–184.

傅军和. (2009). 经典检验 P 值的若干问题. *统计与决策*, (1), 156–157.

傅莺莺, 田振坤, 李裕梅. (2019). 方差分析的回归解读与假设检验. *统计与决策, 35*(8), 77–80.

甘伦知. (2011). 假设检验中控制第二类错误的探讨. *统计与决策*, (22), 35–37.

耿修林. (2008). 方差推断时样本容量的确定. *统计与决策*, (16), 23–25.

龚凤乾. (2003). 统计检验: 实证会计研究方法的核心. *现代财经-天津财经学院学报, 23*(2), 48–51.

郭宝才, 孙利荣. (2010). 关于假设检验中的几个问题的探讨. *统计与决策*, (6), 10–11.

郭璐. (2016). 体育科学研究中统计应用的 7 个误区. *北京体育大学学报, 39*(5), 132–136.

郭文. (2012). 两类错误条件下方差检验中样本容量的确定. *统计与决策*, (9), 12–14.

韩兆洲, 魏章进. (2005). 假设检验的一个常见误区. *统计与信息论坛, 20*(1), 9–11.

郝丽, 刘乐平, 申亚飞. (2016). 统计显著性: 一个被误读的 P 值——基于美国统计学会的声明. *统计与信息论坛, 31*(12), 3–10.

何春. (2011). 正态总体均值与标准差比在序约束下的假设检验. *统计与决策*, (16), 15–16.

贺江南, 张新佶, 谢之辉, 吴美京, 贺佳. (2009). 正态分布资料等效性评价的传统假设检验方法与贝叶斯方法比较. *中国卫生统计, 26*(4), 422–425.

何平平. (2004). 置信区间与假设检验关系中的一个误区. *数理统计与管理*, (4), 77–80.

贺文武. (2004). 浅议零假设及再检验. *统计与决策*, (1), 121–122.

何晓东. (2004). 数据何以“起死回生”——谈外语科研中的显著性检验. *山东外语教学*, (2), 62–64.

侯紫燕, 廖靖宇. (2007). 重复测量试验模型参数似然比检验及其功效分析. *应用概率统计, 23*(1), 68–76.

侯紫燕, 原新凤. (2007). 一类多元重复测量模型参数的似然比检验及其功效分析. *系统科学与数学, 27*(4), 544–554.

胡传鹏, 孔祥祯, Wagenmakers, E.-J., Ly, A., 彭凯平. (2018). 贝叶斯因子及其在 JASP 中的实现. *心理科学进展, 26*(6), 951–965.

胡传鹏, 王非, 过继成思, 宋梦迪, 隋洁, 彭凯平. (2016). 心理学研究中的可重复性问题: 从危机到契机. *心理科学进展, 24*(9), 1504–1518.

胡春健. (2013). 小样本下 Kendall τ 相关系数的显著性检验. *控制工程, 20*(6), 1195–1197.

扈慧敏, 徐兴忠. (2007). 双因素方差分析模型中的广义 p-值. *北京理工大学学报, 27*(9), 843–846.

胡思贵, 王红蕾. (2019). 计数型最小样本量截尾值的序贯检验. *中国科学: 数学, 49*(6), 931–942.

胡竹菁. (2010). 平均数差异显著性检验统计检验力和效果大小的估计原理与方法. *心理学探新, 30*(1), 68–73.

胡竹菁, 戴海琦. (2011). 方差分析的统计检验力和效果大小的常用方法比较. *心理学探新, 31*(3), 254–259.

胡竹菁, 戴海琦. (2017). 心理学实验研究的效果大小. *心理学探新, 37*(1), 70–77.

黄嘉佑. (2005). 气象中使用统计检验的几个问题. *气象,* (7), 3–5.

黄闪闪, 高瑞华. (2014). 逻辑与方法论:贝叶斯统计推理探索的双重视角. *统计与决策,* (15), 4–8.

黄秀秀, 张崇岐. (2014). 混料模型的假设检验. *数理统计与管理, 33*(4), 620–627.

纪竹荪. (2003). 假设检验与区间估计的关系及应用. *统计与决策,* (3), 79–80.

贾旭山, 金振中. (2008). 二项分布贝叶斯假设检验方法. *现代防御技术, 36*(5), 37–40.

贾旭山, 金振中. (2012). Bayes 假设检验及样本数量问题研究. *现代防御技术, 40*(4), 67–70.

江海峰. (2009). 双总体假设检验的另类区间估计法及其 MCS 研究. *统计与决策,* (17), 18–21.

江梅. (2010). 总体相关系数比较的假设检验. *中国卫生统计, 27*(1), 83–87.

江忠伟, 郭新颖. (2018). 基于最小广义特征值的两因素多元方差分析检验法则探讨. *统计与决策, 34*(24), 82–85.

焦璨, 张敏强. (2014). 迷失的边界: 心理学虚无假设检验方法探究. *中国社会科学,* (2), 148–207.

金炳陶, 马承霈. (1992). 检验功效的计算及其应用. *工科数学,* (2), 45–47.

金辉, 邹莉玲. (2017). 假设检验和 P 值的再认识. *环境与职业医学, 34*(2), 95–98.

金晓峰. (2004). 体育统计假设检验中几个问题的探讨. *北京体育大学学报, 27*(9), 1221–1222.

李楚进, 万建平. (2015). 统计检验的发展与应用. *统计与决策,* (23), 2.

李海峰, 姜小峰. (2014). 正确理解和运用比值比反映的效应量. *中国神经免疫学和神经病学杂志, 21*(5), 381.

李康. (2005). 第二讲 数据假设检验的思想与方法. *中国地方病学杂志, 24*(2), 118–119.

李克均, 时松和, 施学忠, 胡东生. (2008). 对应分析应用中的假设检验问题. *中国卫生统计, 25*(2), 199–203.

李路路, 侯艳, 吴莹, 李康. (2014). 设有安慰剂组的多臂非劣效临床试验定量数据的假设检验方法. *中国卫生统计, 31*(6), 1093–1095.

李荣华, 徐九韵. (2001). 多维正态总体零均值的假设检验. *石油大学学报(自然科学版), 25*(6), 112–113.

李胜联, 荀鹏程, 欧超燕. (2006). 变异的假设检验及其应用. *中国卫生统计, 23*(6), 560–561.

李世明, 刘学贞, 徐迪生. (2004). 运动生物力学研究中统计方法应用的几个问题. *广州体育学院学报, 24*(1), 39–41.

李文华, 雷金星. (2005). 假设检验中两类错误的成因、发生概率及其相关问题——以单个总体均值检验为例. *统计与决策,* (4), 117–119.

李勇. (2011). 随机信息中正态均值的灰色统计假设检验判定. *统计与决策,* (22), 29–30.

李勇. (2012). 方差未知的灰色统计假设检验及应用. *统计与决策,* (9), 74–76.

李勇. (2016). 基于两正态均值的灰色统计假设检验研究. *统计与决策,* (1), 19–21.

林丹明, 李炜文, 梁强. (2008). 我国管理学研究的统计功效分析. *中大管理研究, 3*(4), 84–102.

林路, 张润楚. (2001). 假设检验的相对稳定性. *应用数学学报, 24*(4), 616–622.

林晓辉. (2001). 异方差且未知情况下两正态总体等均值检验的贝叶斯观点统计量. *统计与信息论坛, 16*(4), 17–26.

林晓辉. (2006a). 贝叶斯统计学假设检验的一种新方法. *统计与决策,* (16), 9–11.

林晓辉. (2006b). 论模糊数学在假设检验中的应用. *统计与信息论坛, 21*(4), 25–31.

刘红云, 袁克海, 甘凯宇. (2021). 有中介的调节模型的拓展及其效应量. *心理学报, 53*(3), 322–338.

刘佳, 霍涌泉, 陈文博, 解诗薇, 王静. (2018). 心理学研究的可重复性"危机": 一些积极应对策略. *心理学探新, 38*(1), 86–90.

刘媚. (2011). 混合双参数 Pareto 分布的假设检验. *统计与决策,* (2), 34–35.

刘铁川, 王闪闪, 桂雅立. (2019). 方差分析效果大小报告的新指标. *心理学探新, 39*(3), 238–243.

刘玉秀, 徐晓莉, 郑均. (2008). 配对二项数据等效性/非劣效性评价的样本含量估计和假设检验. *中国临床药理学与治疗学, 13*(3), 299–302.

刘韵源, 刘嘉, 陈元立, 周家丽. (2001). 糊状态风险分析的广义 Logistic 回归理论与应用(7)—病例对照研究设计中样本大小与统计功效的估计. *中国公共卫生, 17*(2), 22–23.

刘遵雄. (2018). 类别数据拟合优度检验功效模拟. *统计与决策, 34*(24), 86–87.

刘遵雄, 田珊珊. (2017). 多重假设检验中参数估计问题研究. *统计与决策,* (5), 23–26.

龙兵. (2014). 两参数 Lomax 分布中参数的区间估计和假设检验. *江西师范大学学报(自然科学版), 38*(2), 176–179.

卢谢峰, 唐源鸿, 曾凡梅. (2011). 效应量: 估计、报告和解释. *心理学探新, 31*(3), 260–264.

骆大森. (2017). 心理学可重复性危机两种根源的评估. *心理与行为研究, 15*(5), 577–586.

罗荣华, 吴锟. (2014). 假设检验的一种新思维. *统计与决策,* (8), 23–25.

罗晓娟. (2011). 对假设检验方法的改进. *统计与决策,* (15), 157–158.

吕佳, 乔克林. (2010). 浅谈假设检验中的 P-值. *科学技术与工程, 10*(34), 8494–8496.

吕小康. (2012). Fisher 与 Neyman-Pearson 的分歧与心理统计中的假设检验争议. *心理科学, 35*(6), 1502–1506.

吕小康. (2014). 从工具到范式: 假设检验争议的知识社会学反思. *社会, 34*(6), 216–236.

牟唯嫣, 熊世峰. (2009). 正态密度比的假设检验. *应用概率统计, 25*(6), 632–640.

聂丹丹, 王浩, 罗蓉. (2016). 可重复性: 心理学研究不可忽视的实践. *中国临床心理学杂志, 24*(4), 618–622.
牛莉. (2005). 总体参数单侧检验时如何提出假设 H. *东北林业大学学报, 33*(3), 87–88.
潘高田, 王精业, 杨瑞平. (2001). 小样本离散型多总体和统计量检验法. *系统仿真学报, 13*(2), 182–183.
彭玉兵. (2010). 假设检验中边界样本点的决策. *南昌大学学报(理科版), 34*(4), 346–352.
蒲显伟. (2016). 定量数据分析效应值: 意义、计算与解释. *心理学探新, 36*(1), 64–69.
钱俊, 陈平雁. (2005). 假设检验中计算观察检验效能的意义的探讨. *中国卫生统计, 22*(3), 133–137.
丘甜, 华伟平, 李新光. (2017). 双幂变换下正态线性回归模型参数的假设检验. *统计与决策,* (2), 22–24.
权朝鲁. (2003). 效果量的意义及测定方法. *心理学探新, 23*(2), 39–44.
沈光辉, 范涌峰, 陈婷. (2019). 教育研究中的 P 值使用: 问题及对策——兼谈效应量的使用. *数学教育学报, 28*(4), 92–98.
施能, 章爱国, 余锦华. (2009). 气象学中使用统计检验的几个重要注记. *气象科学, 29*(5), 670–673.
宋立新, 张平. (2009). K 个单参数指数总体相等的假设检验. *东北师大学报(自然科学版), 41*(2), 50–52.
宋爽, 曹一鸣. (2019). 如何正确解读假设检验结果——兼谈数学教育研究中 p 值误用问题. *数学通报, 58*(7), 14–27.
孙红卫, 董兆举, 赵拥军. (2012). 对统计假设检验的误解与误用. *中国卫生统计, 29*(1), 147–150.
孙晓峰, 赵喜春. (2003). 二项分布假设检验平均试验数的确定及其应用研究. *战术导弹技术,* (3), 53–61.
唐宝珍. (2004). 对区间估计和总体参数假设检验思想一致性的思考. *统计与决策,* (2), 125–126.
田庆丰, 张功员. (2002). 医学论文中定量资料假设检验方法常见错误分析. *郑州大学学报(医学版), 37*(1), 70–73.
田晓明, 傅珏生. (2005). 多元总体均值差异显著性检验的研究. *心理科学, 28*(1), 163–165.
万树文, 方芳. (2012). 关于两总体均值差的一种半参数假设检验方法. *中国科学:数学, 42*(7), 671–679.
万霞, 李赞华, 刘建平. (2007). 临床研究中的样本量估算: (1)临床试验. *中医杂志, 48*(6), 504–507.
万霞, 刘建平. (2007). 临床研究中的样本量估算: (2)观察性研究. *中医杂志, 48*(7), 599–601.
王慧, 高雪, 虞明星, 王彤. (2019). logistic 回归中一类基于 Wald 检验的样本量和功效估计. *中国卫生统计, 36*(4), 613–619.
王静, 胡镜清. (2011). 对临床试验中显著性检验、区间检验及置信区间检验之间关系一致性的认识. *中国临床药理学与治疗学, 16*(3), 281–286.
王珺, 宋琼雅, 许岳培, 贾彬彬, 胡传鹏. (2019). 效应量置信区间的原理及其实现. *心理技术与应用, 7*(5), 284–296.
王石青, 史慧娟. (2007). 方差分类模型的假设检验. *河南师范大学学报(自然科学版), 35*(4), 171–172.
王伟. (2004). 医学科研论文中常见的统计学应用错误分析. *中国现代神经疾病杂志, 4*(5), 335–336.
王晓波, 李会琼. (2017). 大维数据中,协方差矩阵等于某个矩阵的假设检验. *云南大学学报(自然科学版), 39*(S1), 24–35.
王潇文, 吕艳. (2020). 一类随机偏微分方程极大似然估计的假设检验. *山东大学学报(理学版), 55*(6), 17–22.
王雪琴. (2010). 关于均值单边检验的局限性. *科学技术与工程, 10*(19), 4740–4743.
王雅慧, 曹春正. (2016). 含方程误差的重复测量误差模型参数的假设检验. *统计与决策,* (4), 16–20.
王雅玲. (2006). 假设检验中无差别区域问题的讨论. *北京工商大学学报(自然科学版), 24*(3), 63–65.
王阳, 温忠麟, 付媛姝. (2020). 等效性检验——结构方程模型评价和测量不变性分析的新视角. *心理科学进展, 28*(11), 1961–1969.
魏杰, 吴学森. (2006). 变量变换对假设检验效能影响的研究. *中国卫生统计, 23*(3), 212–214.
魏艳华, 王丙参, 邢永忠. (2018). 基于蒙特卡洛方法的假设检验问题探讨. *统计与决策, 34*(24), 75–78.
温煦. (2011). 效应量: 体育科研中不应忽略的统计量. *中国体育科技, 47*(3), 142–145.
温忠麟. (2016). *心理与教育统计(第二版)*. 广州: 广东高等教育出版社.
温忠麟, 范息涛, 叶宝娟, 陈宇帅. (2016). 从效应量应有的性质看中介效应量的合理性. *心理学报, 48*(4), 435–443.
温忠麟, 方杰, 沈嘉琦, 谭倚天, 李定欣, 马益铭. (2021). 新世纪 20 年国内心理统计方法研究回顾. *心理科学进展, 29*(8). 1331–1344.
温忠麟, 吴艳. (2010). 屡遭误用和错批的心理统计. *华南师范大学学报(社会科学版)* , (1), 47–54.
温忠麟, 叶宝娟. (2014). 有调节的中介模型检验方法: 竞争还是替补? *心理学报, 46*(5), 714–726.
吴迪, 孙锦峰, 冯丽云. (2007). 假设检验时检验功效的 SAS 实现. *郑州大学学报(医学版)*, *42*(6), 1190–1192.
吴凡, 顾全, 施壮华, 高在峰, 沈模卫. (2018). 跳出传统假设检验方法的陷阱——贝叶斯因子在心理学研究领域的应用. *应用心理学, 24*(3), 195–202.
吴启富, 张玉春. (2012). 统计假设检验中小概率原理的辨析. *统计与决策,* (17), 70–71.
吴艳, 温忠麟. (2011). 与零假设检验有关的统计分析流程. *心理科学, 34*(1), 230–234.
夏佩伦, 李本昌, 李博. (2015). 假设检验在军事工程应用中的若干问题. *火力与指挥控制, 40*(3), 100–103.
夏新涛, 王中宇. (2006). 非统计假设检验原理及其应用. *计量学报, 27*(2), 190–195.
谢田法, 吴启光. (2011). 指数族下参数双边检验的 p–值. *系统科学与数学, 31*(1), 92–104.
谢忠秋. (2018). 权数可靠性的假设检验探讨. *统计与决策, 34*(23), 78–80.
辛自强. (2010). 有关心理统计的三个疑问. *华南师范大学*

学报(社会科学版), (1), 39–46.
徐浪, 马丹. (2001). 假设检验中原假设的确定与 α 控制. *统计与决策,* (12), 14.
徐鹏, 汪卢俊, 严子淳. (2016). 带有随机趋势项的二元选择模型显著性检验研究(英文). *应用概率统计, 32*(3), 301–312.
徐天河, 杨元喜. (2001). 坐标转换模型尺度参数的假设检验. *武汉大学学报(信息科学版), 26*(1), 70–74.
许岳培, 陆春雷, 王珺, 宋琼雅, 贾彬彬, 胡传鹏. (印刷中). 评估零效应的三种统计方法. *应用心理学.*
续志琦, 辛自强. (2018). 单被试实验的统计分析: 非重叠法效果量估计. *心理技术与应用, 6*(2), 89–99.
苟鹏程, 赵杨, 易洪刚, 柏建岭, 于浩, 陈峰. (2006). Permutation Test 在假设检验中的应用. *数理统计与管理, 25*(5), 616–621.
颜杰, 李彩霞, 方积乾, 丁守銮. (2004). 完全随机设计两组 t 检验与秩和检验的功效比较. *中国卫生统计, 21*(1), 12–15.
杨桂元, 刘德志. (2012). 参数假设检验中的若干基本问题研究. *统计与决策,* (24), 13–15.
杨少华, 杨林涛. (2009). 参数假设检验中原假设与备择假设的交换问题. *统计与决策,* (5), 148–149.
姚晨. (2007). 医学研究结论的统计学推断. *北京大学学报(医学版), 39*(2), 213–217.
姚菊香, 王盘兴, 鲍学俊, 卢楚翰. (2007). 相关系数显著性检验的几何意义. *南京气象学院学报, 30*(4), 566–570.
姚嵩坡, 刘盛元, 王滨有. (2010). 假设检验中检验效能的计算及 SAS 实现. *中国卫生统计, 27*(4), 434–436.
尹玉良, 赵俊龙, 徐兴忠. (2011). 正态模型下单边假设检验问题中频率与贝叶斯证据的一致性. *北京理工大学学报, 31*(8), 1001–1004.
余红梅. (2017). 解析美国统计学会关于统计学检验和 P 值的声明. *中国卫生统计, 34*(1), 173–176.
于莉莉, 夏结来, 陈启光, 姚晨. (2005). 显著性检验与等效性检验的区别与联系. *中国卫生统计, 22*(1), 38–39.
曾艳, 李桂花, 庄刘. (2011). 完全随机设计两样本的 Wilcoxon 检验与 K-S 检验功效比较. *中国卫生统计, 28*(4), 372–374.
张超, 胡军, 陈平雁. (2008). 完全随机设计两样本比较的非参数方法的检验功效比较. *中国卫生统计, 25*(3), 230–235.
张凤宽. (2012). 最大熵原理与假设检验方法探讨. *统计与决策,* (15), 10–13.
张弓, 肖景榕. (2006). 正确理解生物统计学的 P 值. *现代肿瘤医学, 14*(1), 102.
张功员. (2002). 医学论文中定性资料假设检验方法的常见错误分析. *编辑学报, 14*(3), 184–186.
张厚粲, 徐建平. (2015). *现代心理与教育统计学.* 北京: 北京师范大学出版社.
张建侠, 鞠银. (2012). 假设检验功效的蒙特卡罗模拟. *统计与决策,* (4), 83–84.
张静. (2012). 贝叶斯假设检验与经典假设检验的对比研究. *统计与决策,* (9), 36–37.
张力为, 祁国鹰. (1998). 效果量: 运动心理学研究应予重视的数据分析指标. *北京体育大学学报,* (01), 13–18.
张凌翔. (2006). 对假设检验中几个问题的思考——兼与韩兆洲、魏章进商榷. *统计与决策,* (6), 32–34.
张晓敏. (2008). 一类马氏样本下假设检验问题错误概率的估计. *应用数学, 21*(1), 180–185.
张应应, 魏毅. (2014). R 函数实现正态总体均值、方差的区间估计及假设检验的设计. *统计与决策,* (9), 74–77.
赵礼, 王晖. (2019). 统计检验力的分析流程与多层模型示例. *心理技术与应用, 7*(5), 276–283.
赵盼, 宋学力. (2016). 泊松分布参数的序贯概率比检验. *统计与决策,* (14), 63–65.
郑发美. (2009). 两均匀分布区间长度比的置信区间与假设检验. *统计与决策,* (22), 152–153.
郑昊敏, 温忠麟, 吴艳. (2011). 心理学常用效应量的选用与分析. *心理科学进展, 19*(12), 1868–1878.
郑文瑞, 丁栋全. (2007). 多元模糊数据的假设检验方法. *模糊系统与数学, 21*(6), 123–127.
钟路. (2004). 对参数单尾假设检验中存在的问题的探讨. *统计与决策,* (11), 27–28.
仲晓波. (2010a). 零假设检验和元分析之间的逻辑连贯性. *心理科学, 33*(6), 1477–1480.
仲晓波. (2010b). 心理学研究中应该怎样报告实验的结果? *心理学探新, 30*(5), 62–65.
仲晓波. (2015). 心理学实验的可重复性. *心理科学, 38*(4), 807–812.
仲晓波. (2016). 关于假设检验的争议: 问题的澄清与解决. *心理科学进展, 24*(10), 1670–1676.
仲晓波, 黄希尧, 万荣根. (2008). 心理学中对假设检验一些批评的分析. *心理科学, 31*(4), 1010–1013.
Anderson, S. F. (2020). Misinterpreting p: The discrepancy between p values and the probability the null hypothesis is true, the influence of multiple testing, and implications for the replication crisis. *Psychological Methods, 25*(5), 596–609.
Champoux, J. E., & Peters, W. S. (1987). Form, effect size and power in moderated regression analysis. *Journal of Occupational Psychology*, *60*(3), 243–255.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
Liu, H., & Yuan, K-H. (2021). New measures of effect size in moderation analysis. *Psychological Methods*, *26*(6), 680–700. https://doi.org/10.1037/met0000371
Liu, H., Yuan, K.-H., & Wen. Z. (in press). Two-level moderated mediation models with single level data and new measures of effect sizes. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01578-6
Lu, Y., & Belitskaya-Levy, I. (2015). p 值之争(英文). *上海精神医学, 27*(6), 381–385.
Miller, J., & Ulrich, R. (2021). A simple, general, and efficient method for sequential hypothesis testing: The independent

segments procedure. *Psychological Methods*, *26*(4), 486–497.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), acc4716.

Rights, J. D., & Sterba, S. K. (2018). A framework of R-squared measures for single-level and multilevel regression mixture models. *Psychological Methods*, *23*(3), 434–457.

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, *24*(3), 309–338.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science, 9*(3), 293–304.

Schnuerch, M., & Erdfelder, E. (2020). Controlling decision errors with minimal costs: The sequential probability ratio t test. *Psychological Methods, 25*(2), 206–226.

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science, 8*(4), 379–386.

Wang, Y. A., & Rhemtulla, M. (2021). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science, 4*(1), 1–17.

Wasserstein, R. L., & Lazar, N. A. (2016). The asa statement on p-values: Context, process, and purpose. *American Statistician, 70*(2), 129–133.

Wen, Z., & Fan, X. (2015). Monotonicity of effect sizes: Questioning kappa-squared as mediation effect size measure. *Psychological Methods, 20*(2), 193–203.

Yuan, K. H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods, 21*(3), 405–426.

Yuan, K. H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(3), 319–330.

Zhang, Z. Y. (2014). Monte Carlo based statistical power analysis for mediation models: Methods and software. *Behavior Research Methods, 46*(4), 1184–1198.

Methodological research on hypothesis test and related issues in China's mainland from 2001 to 2020

WEN Zhonglin[1], XIE Jinyan[1], FANG Jie[2], WANG Yifan[1]

([1] *School of Psychology & Center for Studies of Psychological Application, South China Normal University, Guangzhou* 510631, *China*)

([2] *Institute of New Development & Department of Applied Psychology, Guangdong University of Finance & Economics, Guangzhou* 510320, *China*)

Abstract: In the first two decades of the 21st century, the research of hypothesis test and related topics in China's mainland can be divided into the following categories: Deficiency of null hypothesis significance test, use of *p*-value, repeatability of psychological research, effect size, the power of statistical test, equivalence test, and other research related to hypothesis test. NHST has been developed into a set of procedures as follows. To ensure power of statistical test and save costs, experimental research often needs to do a priori power analysis to estimate the required sample size, while questionnaire studies with more than 160 participants usually does not need to do so for traditional statistical analyses. When the null hypothesis is rejected, a conclusion should be made in combination with an effect size. When the null hypothesis is not rejected, the posterior power of statistical test needs to be reported; if the effect size is medium or large and the power of statistical test is less than 80%, more participants could be added for further analysis, but this process should be disclosed, the final *p*-value should be reported, and the type I error rate should be evaluated.

Key words: hypothesis testing, *p*-value, effect size, power of statistical test, equivalence test